CS 6200 Information Retrieval

Fall 2019

Quiz 7.

Announced Oct 21, 2019
Due Oct 28, 9am Pacific

Succinct answers will be much appreciated.
Open-ended questions may have no 'right' or 'wrong' answer.

1. We looked at various ways to compute the similarity score between a query and a document. The score is what we use to matching documents. Imagine that search engines provide the score for each result, along with the ranking they provide now. Do you think this would be useful? Why, or why not? Explain your answer in 2-3 sentences.

2. Assume we have a collection of N documents, which are numbered 1, 2, 3, …N. What is the idf contribution from a term that occurs:
    a. Exactly twice in each and every document in this whole collection?
    b. Exactly once in each and every odd-numbered document in this collection?
    c. Exactly once in each and every odd-numbered document in this collection and exactly twice in each and every even-numbered document in this collection?

3. Consider the idf definition

    idf of term t = $\log_{10}$ (N/document frequency of t)

    In this equation, we assumed that the log was to the base 10. How will the ranking of documents differ in the vector space tf.idf model of ranking, if you changed the idf log from base 10 to base 8 (octal)?

4. Consider the following table of df and idf for the 4 terms in the table:

Term t	**df_t**	**idf_t**
Nikon	18163	1.65
Canon	6722	2.08
lenses	19240	1.62
tripod	25232	1.5

Consider also the table of term frequencies for three documents D1, D2 and D3:

Term	**D1**	**D2**	**D3**
Nikon	26	5	23
Canon	4	31	0
lenses	0	32	28
tripod	15	0	14

Using the data given, compute the tf-idf weights for the 4 terms above, for each of the three documents.

5. In a few sentences, explain what the tf-idf weighting scheme **ltc.ltn** mean in terms of how you weight the query and documents.